価値関数

⑴ある状態sにいるときに、将来的にどれくらいの報酬が期待できるかを求める。

⑵式：$v_{\pi}(s) = E_{\pi}[G_t|S_t = s]$

$v_{\pi}(s)$	方策πに従った時の、状態sの価値
$E_{\pi}[]$	方策πに従った時の期待値
G_t	時刻tからの割引累計報酬
$S_t = s$	時刻tにおいて状態がsであるという条件

⇓

方策πに従った時の、状態sの価値=方策πに従った時の期待値[時刻tにおいて状態がsであるという条件での時刻tからの割引累計報酬]

(例)
状況
・今、状態s=部屋Aにいる
・行動は「右に進む」「左に進む」のどちらか
・方策πによって、右へ進確率80%、左に進確率20%
・右に進むと、将来の報酬の合計は「50」になる。
・左に進むと、将来の報酬の合計は「30」になる。

価値関数=0.8×50+0.2×30=46
→方策πに従ったときに、状態sにいる価値が46

ベルマン方程式と価値関数

⑴「現在の価値は、即時報酬+次の状態の価値(未来の価値)」
→これをすべての行動と遷移の確率で平均して求める

⑵式：$v_{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s'} P(s'|s,a)[R(s,a,s') + \Upsilon v_{\pi}(s')]$

$v_{\pi}(s)$	方策πに従った時の、状態sの価値
$\sum_{a}$	状態sで選べるすべての行動aに対して合計
$\pi(a\|s)$	方策π：状態sにおいて行動aを選ぶ確率
$\sum_{s'}$	行動aをとったあとに遷移するすべての次の状態s'について合計
$P(s'\|s,a)$	状態遷移確率
$R(s,a,s')$	即時報酬
Υ	割引率
$v_{\pi}(s')$	次の状態s'の価値

⇓

方策πに従った時の状態sの価値=状態sで選べるすべての行動aに対して合計{状態sにおいて行動aを選ぶ確率×行動aをとったあとに遷移するすべての次の状態s'について合計(状態遷移確率×(即時報酬＋割引率＋次の状態s'の価値))}